

The probability of drawing intersections: extending the hypergeometric distribution

Alex T. Kalinka*

Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstr. 108, 01307
Dresden, Germany.

Abstract

The classic hypergeometric distribution describes sampling without replacement from a single urn. Here, I consider a related problem of drawing independently, and without replacement, from two separate urns in which the same n distinct categories of balls exist, but with varying numbers of balls belonging to each category in each urn. The statistic of interest is the size of the intersection of ball categories when we sample independently from each urn. I show that when both urns contain a single ball in each of the categories, the distribution of intersection sizes is hypergeometric. Extending this to the case in which the second urn contains duplicates in $q \leq n$ of its categories, I derive a variant of the hypergeometric distribution and show that it is exact using simulations. I also derive the distribution of absolute pair-wise distances between intersection sizes when sampling separately from two sets of urns and discuss some of the properties and uses of this distribution.

Introduction

The hypergeometric distribution [1] describes the probability of k successes in n draws without replacement from a single population of size N in which reside K possible successes, and is given by

$$P(X = k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}. \quad (1)$$

If, for example, we have 10 red balls in a total population of 20 balls, the above formula gives us the probability of drawing $k \leq 10$ red balls given that we sample $n \leq 20$ balls in total. The distribution has been broadly applied to tests of significance for categorical data in which objects can be classified in two different ways [2–4].

Here, I consider another sampling-without-replacement problem. Imagine instead that we have two separate urns each containing balls that belong to one of n distinct categories of balls. If we draw a balls from the first urn and b balls from the second, what is the probability of finding an intersection of size v in the categories of balls drawn from both urns? Thus, in contrast to the hypergeometric distribution, this problem involves more than 2 categories of objects and independent sampling from two separate urns (Figure 1).

*kalinka@mpi-cbg.de

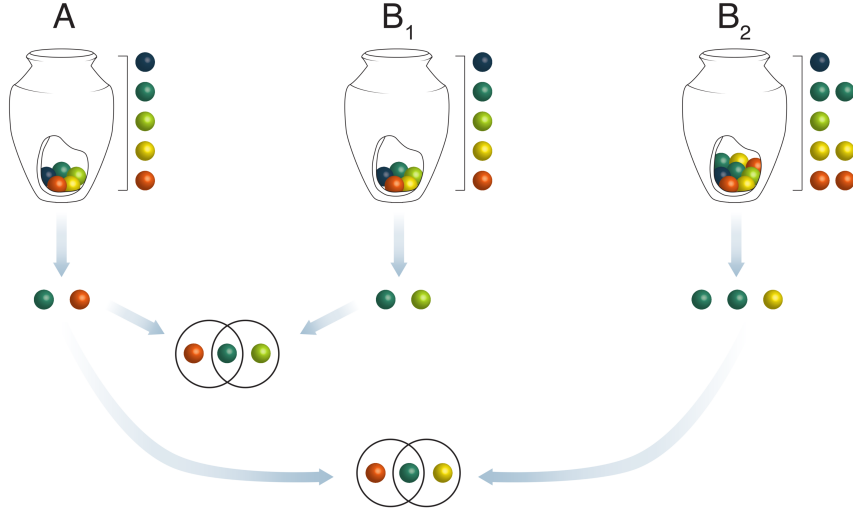


Figure 1. A schematic illustrating the drawing of intersections from urns containing balls belonging to 5 different categories (depicted using different colours). In urns **A** and **B₁** there is exactly 1 ball in each of the categories, whereas in urn **B₂** 3 of the categories contain duplicate members. Although both duplicates of one category are drawn from this urn, the intersection size remains 1.

A real-world example of this type of problem arises in molecular biology. If we know that a set of genes in species 1 shares a particular characteristic (e.g. expression in the same organ) in common with a second set of genes in species 2, then we might ask whether there is a significant enrichment of homologous genes in the intersection of both gene sets (where categories are defined by homology).

Results

Symmetrical, singleton case

First, I consider the simplest scenario in which each urn contains exactly one member in each of the n categories, i.e. a symmetrical, singleton case (corresponding to sampling from urns **A** and **B₁** in Figure 1). We sample $a \leq n$ and $b \leq n$ from each urn respectively and wish to know the probability of drawing intersections of size v where

$$\max(a + b - n, 0) \leq v \leq \min(a, b).$$

To count the number of ways of picking an intersection of size v , it is useful to note that for the first urn there are $\binom{n-v}{a-v}$ ways to draw one particular combination of intersecting categories (e.g. categories 1,2,3 for $v = 3$). However, once we have drawn from the first urn, this leaves $\binom{n-a}{b-v}$ ways of drawing from the second urn to give an intersection size of v for a single, specific combination of categories - the upper index of $(n - a)$ ensures that we do not count intersections of size larger than v . The total number of ways to pick intersections of size v must then be summed over all $\binom{n}{v}$ category combinations:

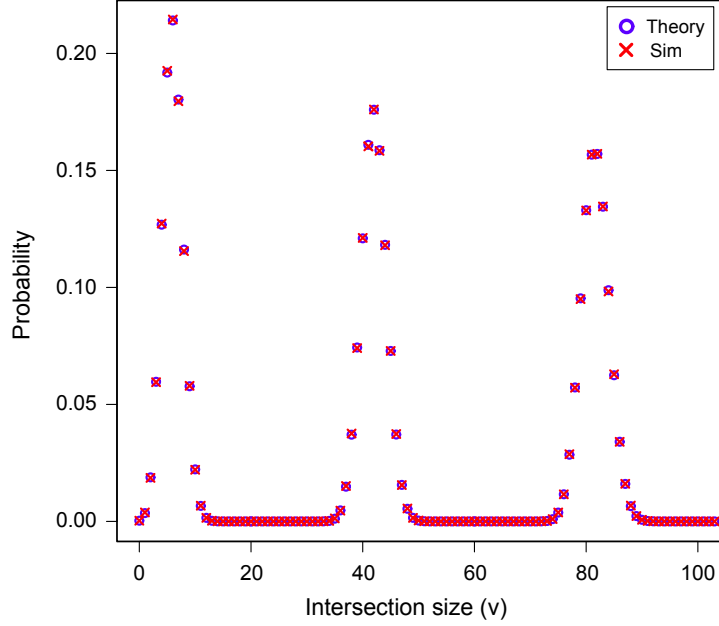


Figure 2. Match between theory and simulation for 3 parameter sets in the symmetrical, singleton case (**A** and **B**₁ in Figure 1). Simulations (Sim) consisted of randomly and independently sampling twice (without replacement) from n distinct categories and recording the size of the intersection each time (repeated 500,000 times for each distribution). From left to right, the parameters were: $n = 100, a = 20, b = 30$; $n = 100, a = 70, b = 60$; $n = 155, a = 110, b = 115$.

$$C_v = \sum_{a-v}^n \binom{n-v}{a-v} \binom{n-a}{b-v} = \binom{n}{v} \binom{n-v}{a-v} \binom{n-a}{b-v}.$$

The probability of picking an intersection of size v is then C_v divided by the total number of ways of picking a and b from n :

$$P(X = v) = \frac{\binom{n}{v} \binom{n-v}{a-v} \binom{n-a}{b-v}}{\binom{n}{a} \binom{n}{b}}. \quad (2)$$

Applying a trinomial revision [5] to the first two binomials in the numerator, the expression can be reduced to

$$P(X = v) = \frac{\binom{n}{a} \binom{a}{v} \binom{n-a}{b-v}}{\binom{n}{a} \binom{n}{b}} = \frac{\binom{a}{v} \binom{n-a}{b-v}}{\binom{n}{b}}, \quad (3)$$

which is the hypergeometric distribution given in Equation 1, and is symmetrical in terms a and b . The symmetry of the problem, in which both urns contain exactly 1 member in each of the n categories, enables this simplification. Simulations show that the distribution is exact (Figure 2).

Asymmetrical, duplicates case

If we allow duplicates in $q \leq n$ of the categories in the second urn (each category can contain 1 or 2 balls but not 0), then the problem becomes asymmetrical (corresponding to **A** and **B₂** in Figure 1). The presence of duplicates in the second urn will reduce the overall chance of drawing an intersection of a certain size because duplicates that are both sampled from the urn can at most contribute an intersection of size 1 (see Figure 1). Thus, it is reasonable to conjecture that the expectation for this distribution will always be less than for the equivalent symmetrical, singleton case:

$$E(X|q \neq 0) < \frac{ab}{n}. \quad (4)$$

Furthermore, we can reasonably suppose that the expression describing this distribution will be a variant of the hypergeometric since we have made only a small modification to the basic problem, and added a single parameter, q . I begin, therefore, by modifying Equation 2 to account for the effects of including q duplicates.

There are three main differences affecting the drawing of both intersecting and non-intersecting categories:

1. When a category is sampled from the first urn (among the $a - v$ non-intersecting categories) for which there is a duplicate pair in the equivalent category in the second urn, this reduces the number of ways we can pick non-intersecting categories from the second urn to ensure an intersection size of v . The number of such draws is indicated by the subscript m .
2. If a category with a duplicate pair is picked in the v intersecting items, this does not reduce the number of ways of picking non-intersecting items from the second urn (since drawing the duplicate member will also produce an intersection of size v), but it removes a duplicate category from the m that could be picked in the non-intersecting set. The number of such draws is indicated by subscript l .
3. Picking l duplicate categories in the v intersecting items increases the number of ways that these items can be drawn, but for each duplicate that is picked in v , one less duplicate is available for the non-intersecting set to be drawn from the second urn. The number of such draws is indicated by the subscript j .

To calculate the probability, we must sum over all of the ways of combining the above events such that they produce intersection sizes of v , which must satisfy

$$\max(a + b - n - q, 0) \leq v \leq \min(a, b).$$

I will move from left to right across the numerators of Equation 2 and describe how each binomial term must be modified. The number of ways of drawing v intersecting categories, $\binom{n}{v}$, must incorporate consideration for the l duplicate categories listed in point 2 above, leading to:

$$\sum_l \sum_j \binom{n-q}{v-l} \binom{q}{l} \binom{l}{j},$$

which counts the total number of ways of picking v with and without l duplicates. The number of ways of picking non-intersecting categories from the first urn for a

single category combination, $\binom{n-v}{a-v}$, must be modified to account for the m duplicate categories listed in point 1 above:

$$\sum_m \sum_l \binom{n-v-q+l}{a-v-m} \binom{q-l}{m},$$

which counts the number of ways of picking $a-v$ non-intersecting categories given that we have sampled both m and l duplicates. Finally, the number of ways of picking non-intersecting items from the second urn, $\binom{n-a}{b-v}$, must be modified to account for a reduction in duplicates that can be drawn to ensure an intersection size of v :

$$\sum_m \sum_j^l \binom{n+q-a-m-j}{b-v},$$

which counts the number of ways of picking $b-v$ non-intersecting categories from the second urn given that we have picked m non-intersecting duplicate equivalents and j intersecting duplicate equivalents from the first urn.

Summing over all the possible combinations of these events then gives us the total number of ways of picking an intersection of size v in the duplicate case (underbraces indicate equivalent expressions in the symmetrical singleton case in Equation 2):

$$C_v^d = \sum_{m=0}^{\beta} \sum_{l=0}^{\gamma} \sum_{j=0}^l \underbrace{\binom{n-q}{v-l} \binom{q}{l} \binom{l}{j}}_{\binom{n}{v}} \underbrace{\binom{q-l}{m} \binom{n-v-q+l}{a-v-m}}_{\binom{n-v}{a-v}} \underbrace{\binom{n+q-a-m-j}{b-v}}_{\binom{n-a}{b-v}},$$

where

$$\beta = \min(a-v, q) \quad \text{and} \quad \gamma = \min(v, q-m).$$

The probability is then C_v^d divided by the total number of ways of picking a and b from both urns:

$$P(X=v) = \frac{\sum_{m=0}^{\alpha} \sum_{l=0}^{\beta} \sum_{j=0}^l \binom{n-q}{v-l} \binom{q}{l} \binom{l}{j} \binom{q-l}{m} \binom{n-v-q+l}{a-v-m} \binom{n+q-a-m-j}{b-v}}{\binom{n}{a} \binom{n+q}{b}} \quad (5)$$

A closed-form expression for the above equation is not forthcoming, however, it is clear that the following identity is true:

$$\sum_{v=0}^{\alpha} C_v^d = \binom{n}{a} \binom{n+q}{b},$$

where $\alpha = \min(a, b)$. Once again, simulations show that the distribution is exact (Figure 3).

The distribution of intersection distances

When drawing intersections from two different distributions (with different parameters, or, for example, with a singleton case and a duplicate case) it might be of interest to

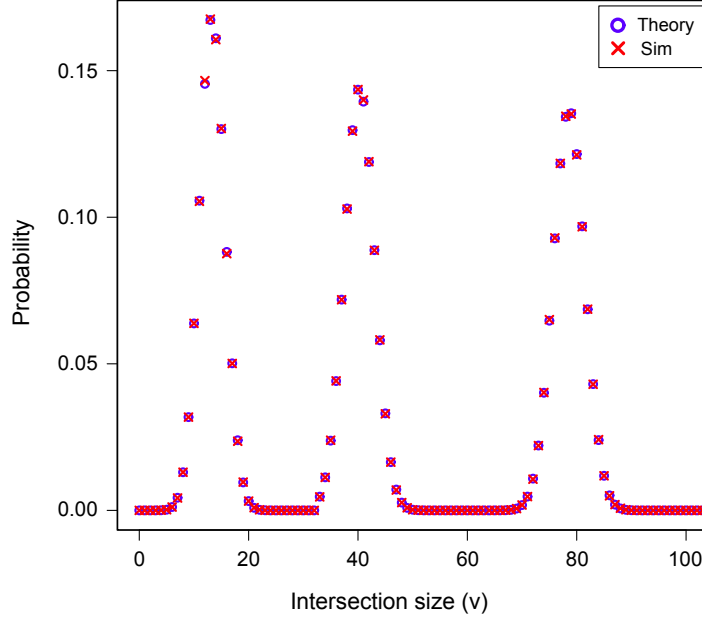


Figure 3. Match between theory and simulation for 3 parameter sets in the asymmetrical, duplicate case (**A** and **B**₂ in Figure 1). Simulations (Sim) consisted of randomly and independently sampling twice (without replacement) from n distinct categories (in which the second set contained q duplicates) and recording the size of the intersection each time (repeated 500,000 times for each distribution). From left to right, the parameters were: $n = 100, a = 35, b = 42, q = 59$; $n = 100, a = 63, b = 79, q = 73$; $n = 130, a = 110, b = 115, q = 47$.

ask whether the absolute distance between their intersection sizes is what would be expected by chance. To calculate the probability of finding an intersection distance of size d , we need to sum over all the ways to produce d when pairing our two distributions:

$$P(X = d) = \sum_{\{v_1, v_2\}_i \in D_d}^{|D_d|} P(v_{1_i} | n_1, a_1, b_1, \dots) \cdot P(v_{2_i} | n_2, a_2, b_2, \dots) \quad (6)$$

where D_d is the set of pairs of intersection sizes, $\{v_1, v_2\}$, with absolute differences of size d . If R and S are the sets of all possible intersection sizes for both distributions, then

$$|D_d| = \begin{cases} |R \cap S|, & \text{if } d = 0 \\ \min(|R|, |S| - d) + \min(|S|, |R| - d), & \text{if } d > 0 \end{cases}$$

where $\min(|R|, |S| - d), \min(|S|, |R| - d) \geq 0$ (i.e. negative values are set to zero). This distribution has a relationship to the intersection distributions that is similar to the relationship between the binomial and Bernoulli distributions.

Testing for significant differences between intersection sizes is likely to be of interest when we want to know if they are behaving differently, i.e. are the intersection

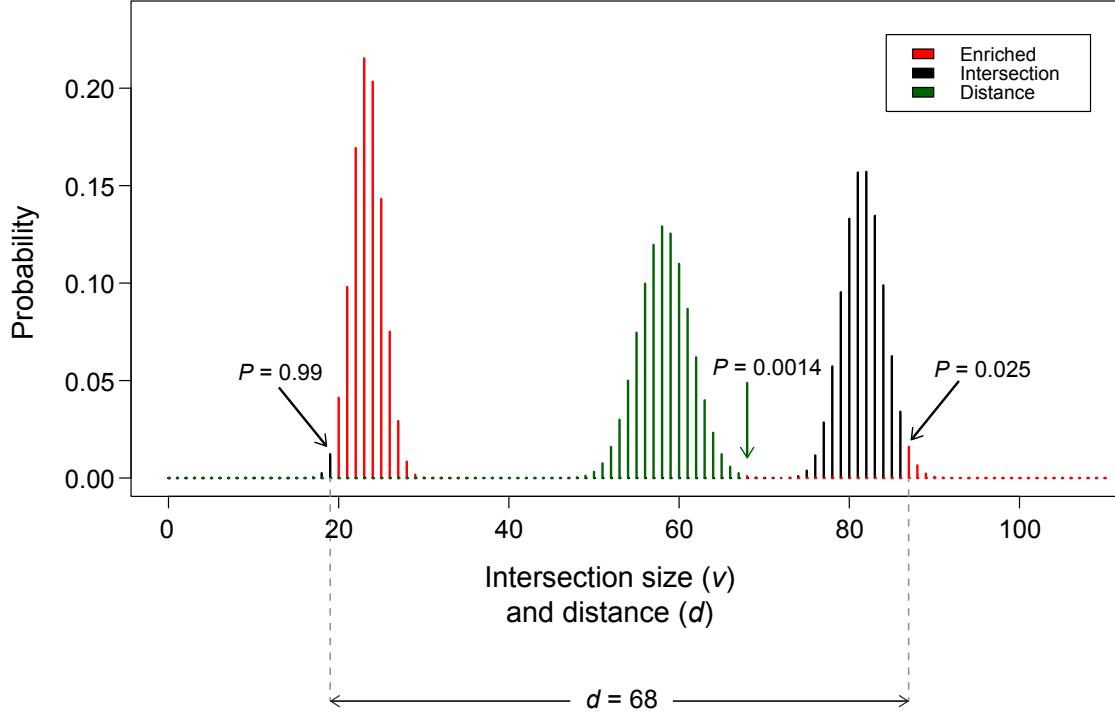


Figure 4. Two intersection distributions (in black and red) and their distance distribution (in dark green and red). One-tailed tests for greater intersection sizes than expected by chance have been applied to both intersection distributions at 19 (left) and 87 (right). This gives a distance of 68, which is greater than expected by chance ($P = 0.0014$) even though significance at the 5% level was marginal for only one of the intersection distributions. Parameter values for the two intersection distributions were, from left to right: $n = 60, a = 40, b = 35$; $n = 155, a = 110, b = 115$.

sizes that we observe falling into opposite tails more than would be expected by chance (Figure 4)?

Related distributions: drawing from a single urn

It is useful to consider the related, though simpler, distribution associated with drawing a balls from n categories with $q \leq n$ duplicates from a single urn. In this case, we are no longer interested in intersection sizes, but rather in the number of distinct categories, c , which are drawn from the single urn. When $q = 0$ then $c = a$ necessarily. Thus, we restrict ourselves to cases where $0 < q \leq n$. The bounds on c are then

$$a - \min\left(\left\lfloor \frac{a}{2} \right\rfloor, q\right) \leq c \leq a$$

where the lower bound is determined by the maximum possible number of duplicate pairs subtracted from a . For any particular value of c , there are always $a - c$ duplicate pairs that must be picked to ensure that there are c distinct categories drawn.

To count the number of ways of drawing c categories, it is useful to first note that there are 3 combinations that need to be counted:

1. The number of ways of picking $a - c$ duplicate pairs, i.e. $\{1,1\}, \{2,2\}$.
2. The number of ways of picking duplicates not picked as pairs from the $q - a + c$ remaining.
3. The number of ways of picking non-duplicates.

Point 1 is simply given by $\binom{q}{a-c}$. For points 2 and 3, we must sum over all the ways of combining duplicates (not picked as pairs) and non-duplicates to give c distinct categories. An important quantity here is the number of non-duplicate pairs (not $\{1,1\}$ or $\{2,2\}$) in a , given by $a - 2(a - c) = 2c - a$. Then

$$\sum_{j=0}^q \binom{q-a+c}{j} \binom{n-a+c-j}{2c-a-j}$$

gives the combined number for points 2 and 3. The probability of picking c distinct categories is then given by

$$P(X = c) = \frac{\binom{q}{a-c} \sum_{j=0}^q \binom{q-a+c}{j} \binom{n-a+c-j}{2c-a-j}}{\binom{n+q}{a}}. \quad (7)$$

Again, a closed-form expression is not easily derived. However, if we focus on the special case when $q = n$, the above expression can be simplified. Substituting $q = n$ and applying a trinomial revision to the binomials within the sum, the numerator can be reduced to

$$\binom{n}{a-c} \binom{n-a+c}{2c-a} \sum_{j=0}^n \binom{2c-a}{j},$$

which in turn simplifies to

$$\binom{n}{c} \binom{c}{a-c} 2^{2c-a}.$$

From left to right, the three terms count the number of ways of picking c distinct categories from n , the number of ways of picking $a - c$ duplicate pairs from c , and the number of ways of picking $2c - a$ duplicates *not* picked as duplicate pairs. The probability of drawing c distinct categories when all categories in the urn contain a duplicate is then given by

$$P(X = c) = \frac{\binom{n}{c} \binom{c}{a-c} 2^{2c-a}}{\binom{2n}{a}}. \quad (8)$$

Deriving the expectation of this distribution requires simplifying

$$\sum_{c=\beta}^a c \binom{n}{c} \binom{c}{a-c} 2^{2c-a} = n 2^{-a} \sum_{c=\beta}^a \binom{n-1}{c-1} \binom{c}{a-c} 2^{2c}$$

where $\beta = a - \lfloor \frac{a}{2} \rfloor$. Again, a closed-form is not apparent. Future work on this distribution and its properties will no doubt help to shed light on the related intersection distributions.

Summary

I have discussed a sampling-without-replacement problem that is closely related to the hypergeometric distribution. The main differences are:

1. Samples are taken independently from two separate urns.
2. More than 2 categories of objects are allowed.
3. The statistic of interest is the size of the intersection of the two samples.

When there is exactly one ball in each of the categories in both urns, I have shown that the distribution is hypergeometric. However, with one small modification to the basic problem - the addition of q duplicate categories in the second urn - the problem becomes much more complex, and the distribution no longer can be expressed in a closed form, though it is clear that this asymmetrical case is a variant of the hypergeometric. Furthermore, I derive the distribution of absolute distances between the intersection sizes of two separate intersection distributions. This distribution has utility when the question of interest is whether two intersection sizes are behaving significantly differently. Finally, I derived a closed-form expression for the distribution of distinct categories sampled from a single urn containing duplicates in all of its categories. This related distribution may aid in the understanding of intersection distributions and their properties, and ultimately is an attempt to work towards a more general description of this broad class of distributions.

Acknowledgements

I gratefully acknowledge Peter Steinbach's assistance with implementing the duplicate case in C++ for large parameter sets, and Iva Kelava for preparing Figure 1.

References

- [1] Huygens C (1657) De Ratiociniis in Ludo Aleae or The Value of all Chances in Games of Fortune.
- [2] Pearson K (1899) On certain properties of the hypergeometrical series, and on the fitting of such series to observation polygons in the theory of chance. Philosophical Magazine 47: 236-246.
- [3] Fisher RA (1922) On the interpretation of χ^2 from contingency tables, and the calculation of p . J Roy Statist Soc 85: 87-94.
- [4] Gonin HT (1936) The use of factorial moments in the treatment of the hypergeometric distribution and in tests for regression. Philosophical Magazine 21: 215-226.
- [5] Graham RL, Knuth DE, Patashnik O (1994) Concrete Mathematics: A Foundation for Computer Science. Addison Wesley, 2nd edition.